Data Model and Management of Plant Genetic Resources

Masamichi UMEHARA*

Abstract

Data management system of the NIAR gene bank was developed on the basis of a relational data model using relational data base management system (RDBMS) which is running under the UNIX operating system. Stored data are classified into three categories: "passport", "stock", "evaluation"data. A total of 170,000 records has been compiled for passport data, along with 300,000 stock data and 100,000 evaluation data, so far.

Data model of plant genetic resources is as follows:

The schema for the passport data management consists of 12 tables, *e. g.* a passport data table, a plant code table, etc.

The schema for the stock data management consists of 14 tables, e. g. tables related to the registration of new accessions, distribution of seeds, multiplication of seeds, germination test of seeds, etc.

The schema for the management of evaluated data consists of more than 600 tables. Evaluation was performed for approximately 110 plants.

As usually data management of one table requires a data management program, the passport data management system and the stock data management system were readily developed, under such concept.

For the management of evaluated data, however, it is impossible to develop 600 programs for the management of 600 tables. Therefore, instead of the development of specific programs for each plant, we developed a data dictionary system relating to the table structure for each plant and also a data management program which covers the whole tables for the plants using the dynamic SQL.

Outline of schema

A data model for the plant genetic resources stored at NIAR has been developed on the basis of a relational data model. This schema consists of about 700 tables, which can be classified into 3 categories (Fig. 1).

In the first category, passport data are the basis of the plant genetic resources database. For the identification of individual genetic resources, the passport data are firstly constructed at the time of collection or reception of the accessions. Every genetic resource has it's own accession number which is never duplicated among the plant genetic resources of MAFF.

In the second category, the stock control data are very important for the management of plant genetic resources, because the distribution of seeds upon request by researchers is a major working. In this management, data such as germination rates, seed weight, addresses in the preservation room are essential. Each plant genetic resources is identified only by the accession number.

In the third category, data of the evaluated characteristics are most useful for users of genetic resources. Table structures of evaluation data may/ should be different for each species. Then, a large

Presented at the 27th International Symposium on "Plant Genetic Resource Management in the Tropics", Tsukuba, Ibaraki, Japan, 25-26 August 1993, held by Tropical Agriculture Research Center (TARC).

* Department of Genetic Resources I, National Institute of Agrobiological Resources, Tsukuba, Ibaraki 305 Japan

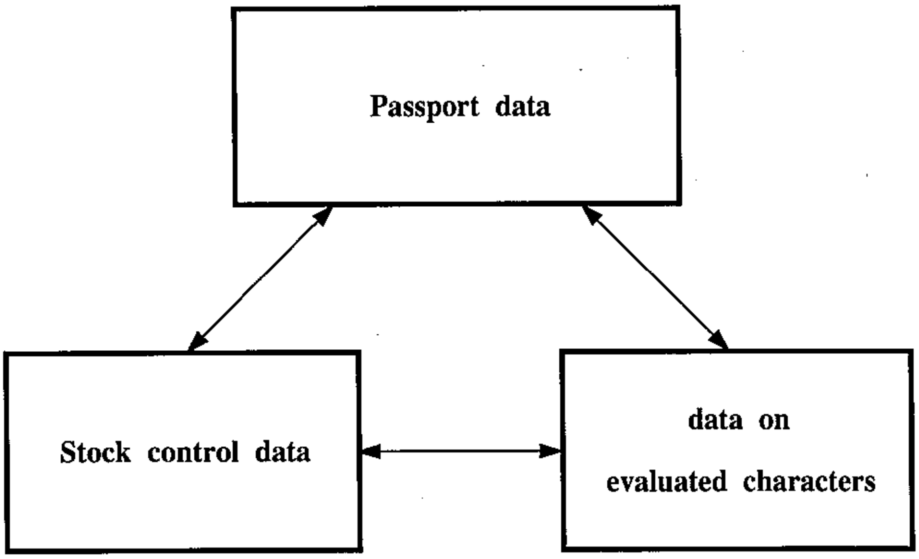


Fig. 1 Data model for plant genetic resources

number of tables, presently over 600 tables, is required. The management of evaluated data is, therefore, the most complex procedure.

Passport data

The basic structure of the passport data is shown in Fig. 2. For the purpose of normalization, every row of passport data contains an accession number, an institute code, a laboratory code, a plant code and an origin code. As a result of normalization, the passport data table refers to 4 foreign tables,: i. e. an institution table, a laboratory table, a plant table and a country table. Passport data are subdivided and stored in 3 tables of the same structure, "pass", "pass_sub" and "pass_wk" based on actual efficiency of data management. Some other additional tables are also used for the data management such as determination of accession number. The accession numbers of passport data are, then, used for the stock control data and for the character evaluation data.

Stock control data

"Address", a table for storage addresses of seed bottles in the preservation room, has columns for accession numbers, addresses, germination rates, amount of seeds, warehousing days, etc.

Two tables are used for the distribution, "request_no" and "request_item". The table "request_no" is used for the attributes of user who requested and "request_item" is used for the accessions requested.

For the multiplication and for the test of germination rates, tables of a similar structure are used.

Fig. 3 shows tables for the distribution.

Evaluation data

All the evaluation data commonly have an accession number, an institution code, a laboratory code in which the genetic resource was evaluated, and the year of evaluation. Other attributes vary depending on the plant groups. All the attributes of characteristics which are evaluated for each plant are included in a data dictionary system. The number of tables for the data of evaluated characteristics are over 600, and those tables are created automatically by a program (Fig. 4).

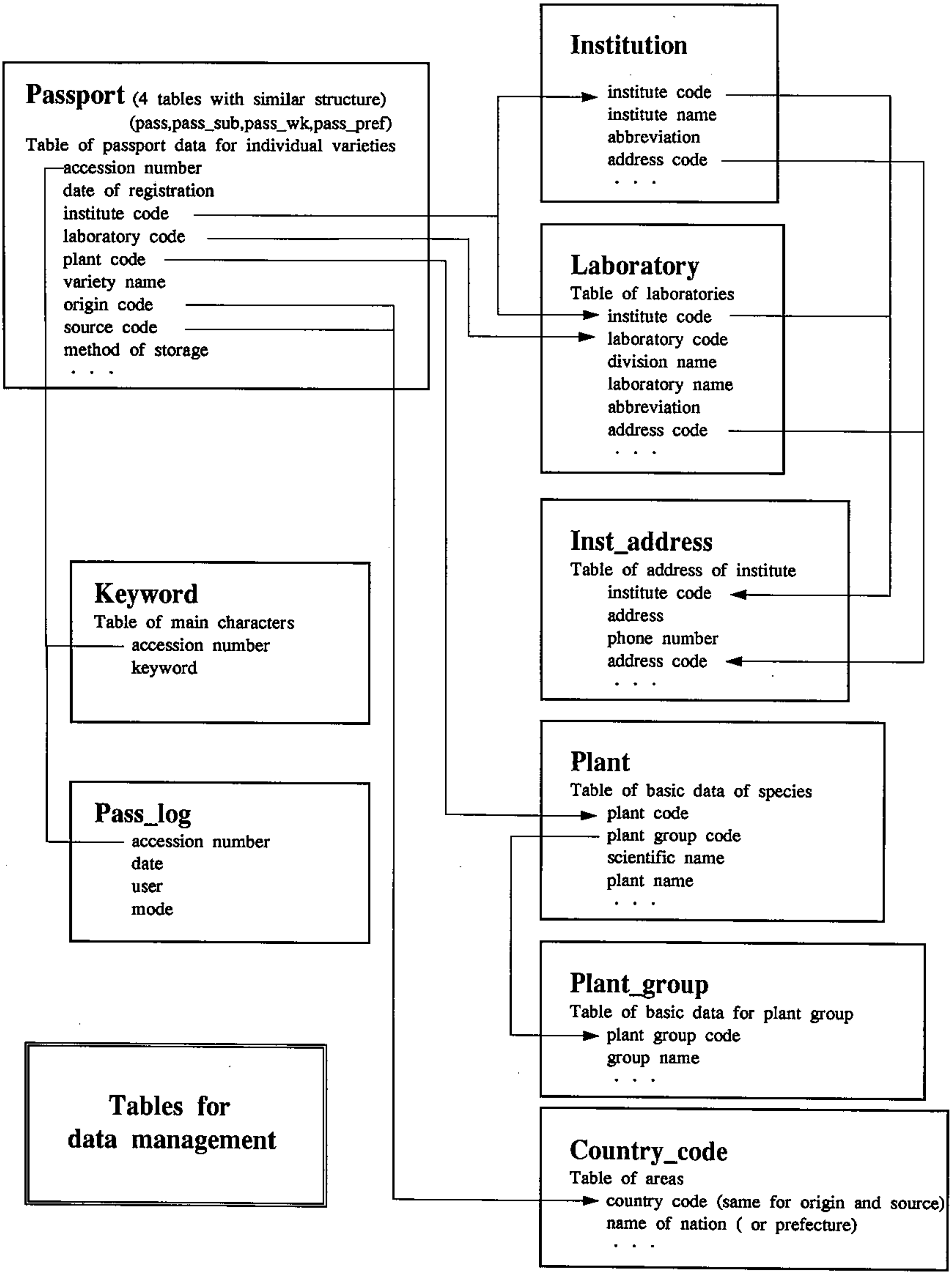


Fig. 2 Passport data

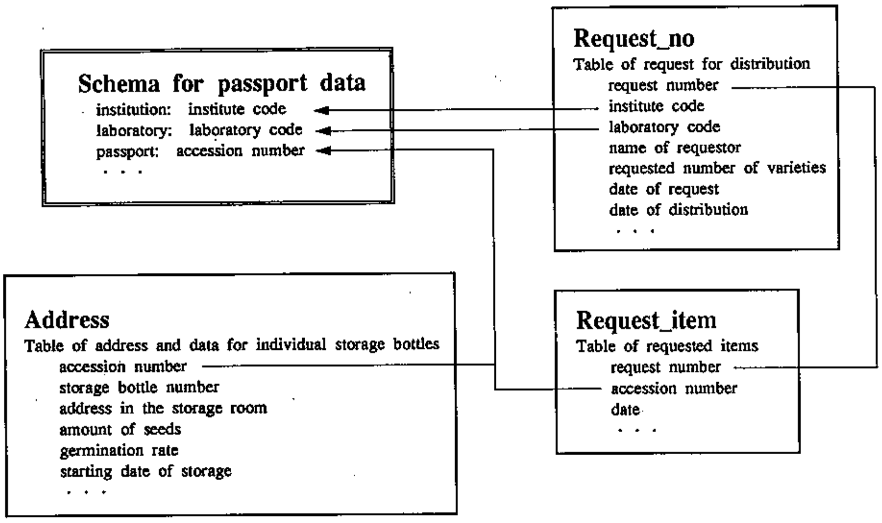


Fig. 3 Tables for the distribution

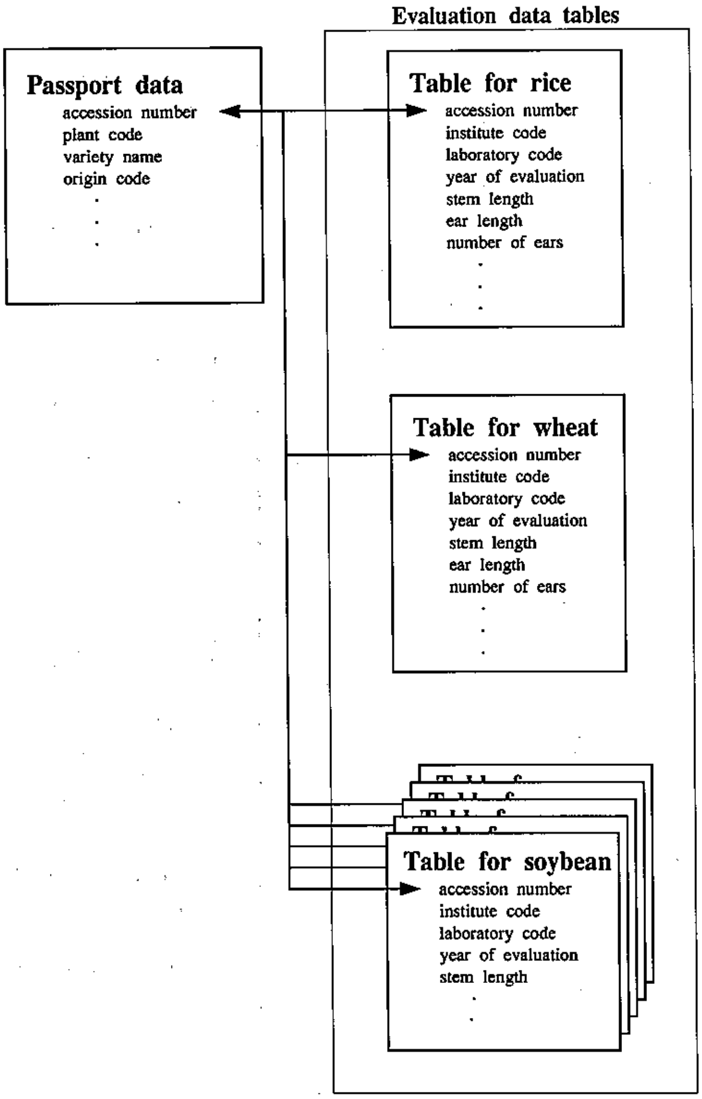


Fig. 4 Evaluation data

Data processor in the genebank, NIAR

Fig. 5 shows the concept of data processing in the genebank. Application programs for data management are executed on a UNIX operating system. In the database server, 10 to 20 processes are normally running for the database access.

Data management

A data management system has been developed for the passport data and the stock control data by an ordinary method.

The stock control management consists of 4 sub-systems, as follows :

1) New seed reception
2) Distribution of genetic resources
3) Multiplication of genetic resource
4) Germination testing

Ability for OLTP (OnLine Transaction Processing) of DBMS is required for the management of the stock control data.

Data management for the evaluated characteristics

By using ordinary methods, the development of a data management system for evaluated characteristics is difficult, because too many tables are required. Instead of using ordinary methods, we introduced a data dictionary system.

In the first step, we developed a management system for the data dictionary, then we input whole data which define the characteristics for the evaluation of each plant.

In the next step, a program which generates SQL statement for creating tables was developed, by us-

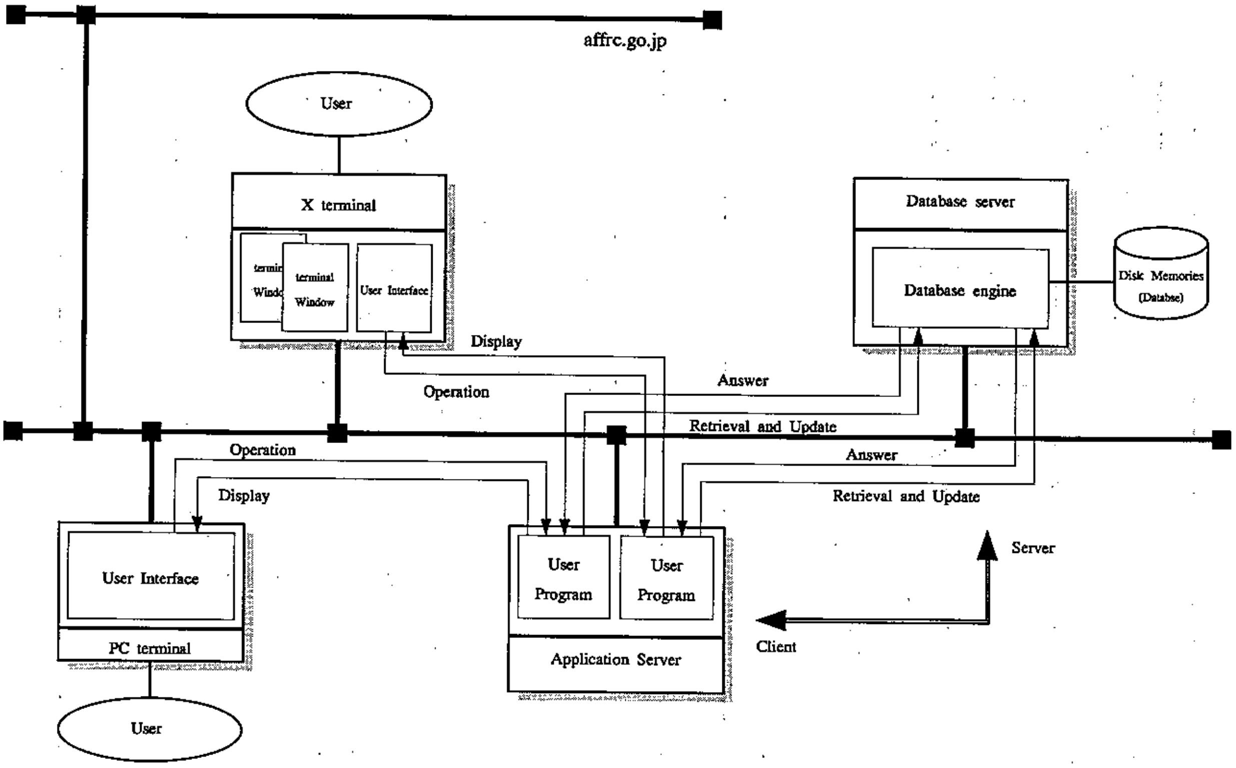


Fig. 5 Outline of data processing in genebank, MAFF

ing the data dictionary. Fig. 6 shows SQL statements generated.

In the last step, a program which can handle the whole evaluation table by using the data dictionary was developed. Operators can select any table of evaluation data with a menu system. After the decision of the plant, this program selects whole definition data from data dictionary for the plant, then allocates necessary memories and manages the data of the table by the dynamic SQL (Fig. 7).

Fig. 8-Fig. 11 show examples of screen images.

```
CREATE TABLE t 02001 h 1 (
    acce_no     char (8) NOT NULL,
    inst_code   char (5) NOT NULL,
    lab_code    char (5) NOT NULL,
    year        char (4) NOT NULL check (year > “1970”),
    h 1_001     char (1) check (h 1_001 between“2”and“8”),
    h 1_002     integer CHECK (h 1_002 > 0 AND h 1_002 < 9999),
    h 1_003     decimal (5, 1) CHECK (h 1_003 > 0 AND hi_003 <999.9),
    h 1_004     char (1) check (h 1_004 in (“0”, “2”, “3”, “4”, “5”, “6”, “7”, “8”)),
    h 1_005     char (1) check (h 1_005 between“1”and“9”),
    h 1_006     char (1) check (h 1_006 between“2”and“8”),
    h 1_007     char (1) check (h 1_007 between“0”and“9”),
    h 1_008     char (5),
    h 1_009     char (5),
    foreign key (inst_code) REFERENCES inst_code,
    foreign key (inst_code, lab_code) REFERENCES lab_code,
    Primary key (acce_no, inst_code, lab_cord, year)
);
```

Fig. 6 Example of SQL statements generated

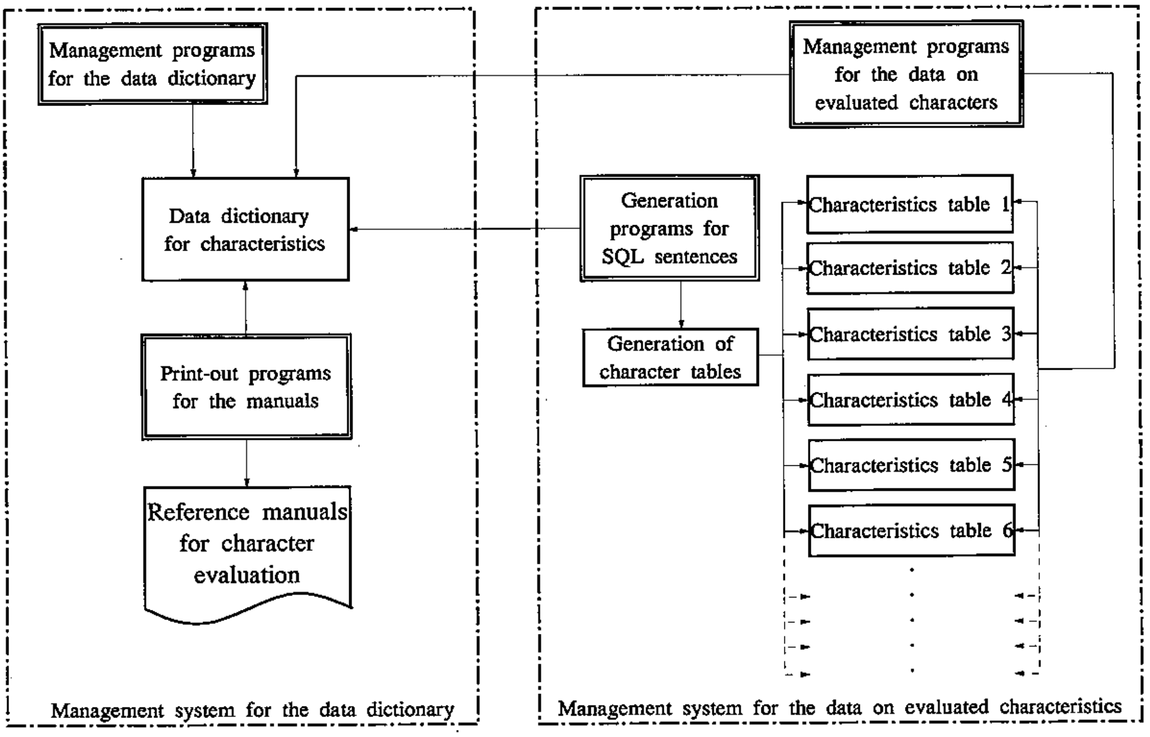
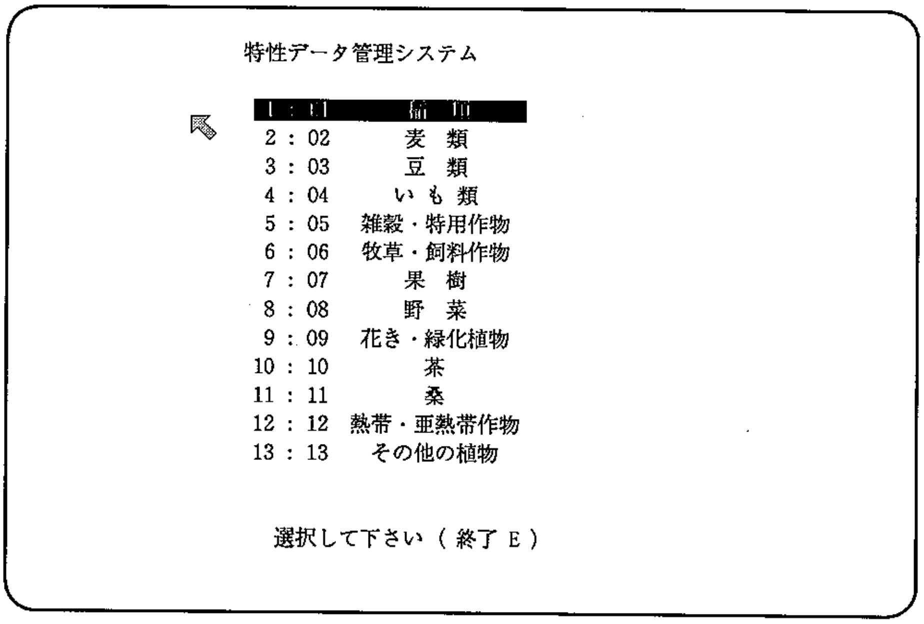

Management programs for the data dictionary
Data dictionary for characteristics
Print-out programs for the manuals
Reference manuals for character evaluation
Management system for the data dictionary
Management programs for the data on evaluated characters
Generation programs for SQL sentences
Generation of character tables
Characteristics table 1
Characteristics table 2
Characteristics table 3
Characteristics table 4
Characteristics table 5
Characteristics table 6
Management system for the data on evaluated characteristics

Fig. 7 Data management for the evaluated characteristics

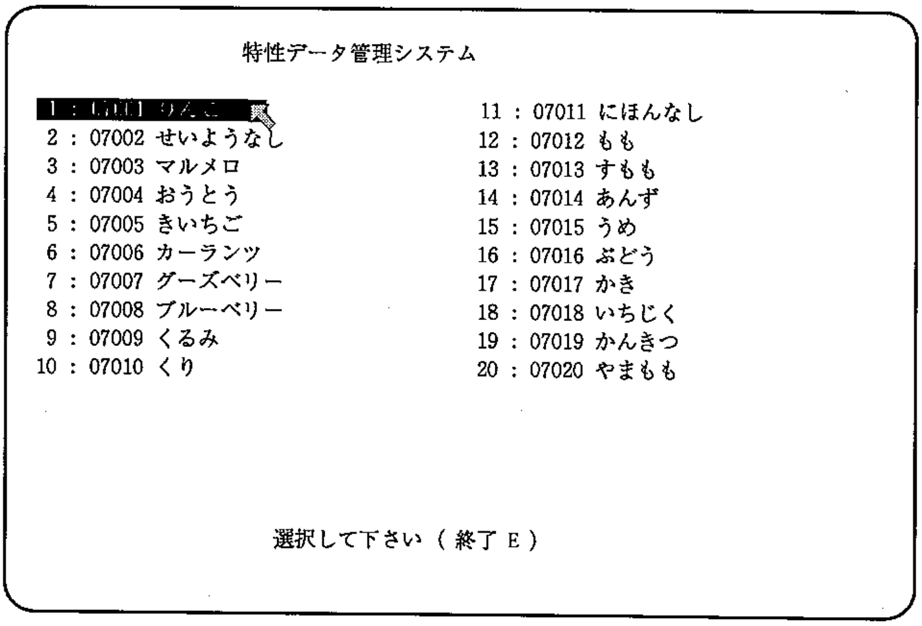


```
特性データ管理システム

 1 : 01       稲  類
 2 : 02       麦  類
 3 : 03       豆  類
 4 : 04      い も 類
 5 : 05    雑穀・特用作物
 6 : 06    牧草・飼料作物
 7 : 07       果  樹
 8 : 08       野  菜
 9 : 09    花き・緑化植物
10 : 10        茶
11 : 11        桑
12 : 12   熱帯・亜熱帯作物
13 : 13     その他の植物

選択して下さい（ 終了 E ）
```

Fig. 8　Top menu

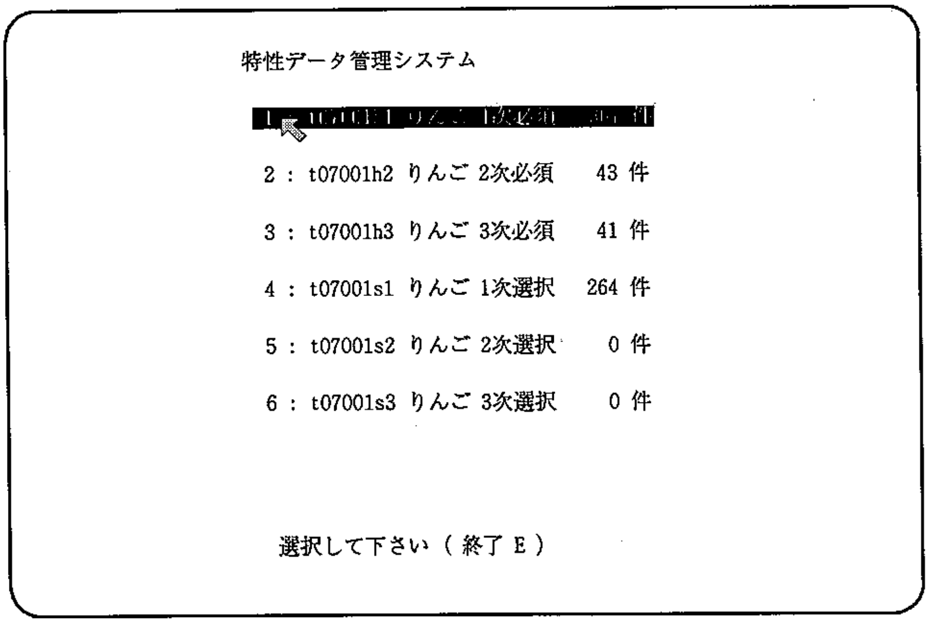


```
特性データ管理システム

 1 : 07001 りんご                11 : 07011 にほんなし
 2 : 07002 せいようなし          12 : 07012 もも
 3 : 07003 マルメロ              13 : 07013 すもも
 4 : 07004 おうとう              14 : 07014 あんず
 5 : 07005 きいちご              15 : 07015 うめ
 6 : 07006 カーランツ            16 : 07016 ぶどう
 7 : 07007 グーズベリー          17 : 07017 かき
 8 : 07008 ブルーベリー          18 : 07018 いちじく
 9 : 07009 くるみ                19 : 07019 かんきつ
10 : 07010 くり                  20 : 07020 やまもも

選択して下さい（ 終了 E ）
```

Fig. 9　Menu for fruits



特性データ管理システム

1 : t07001h1 りんご 1次必須 84 件

2 : t07001h2 りんご 2次必須 43 件

3 : t07001h3 りんご 3次必須 41 件

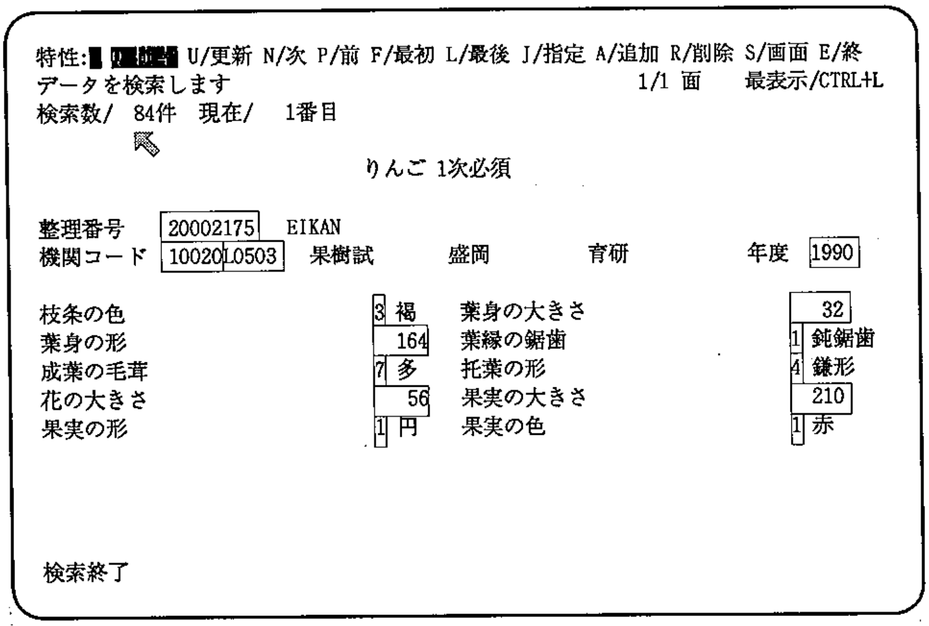
4 : t07001s1 りんご 1次選択 264 件

5 : t07001s2 りんご 2次選択 0 件

6 : t07001s3 りんご 3次選択 0 件

選択して下さい (終了 E)

Fig. 10 Menu for apple



特性:▌ [illegible] U/更新 N/次 P/前 F/最初 L/最後 J/指定 A/追加 R/削除 S/画面 E/終
データを検索します 1/1 面 最表示/CTRL+L
検索数/ 84件 現在/ 1番目

りんご 1次必須

整理番号 20002175 EIKAN
機関コード 10020 10503 果樹試 盛岡 育研 年度 1990

枝条の色	3 褐	葉身の大きさ	32
葉身の形	164	葉縁の鋸歯	1 鈍鋸歯
成葉の毛茸	7 多	托葉の形	4 鎌形
花の大きさ	56	果実の大きさ	210
果実の形	1 円	果実の色	1 赤

検索終了

Fig. 11 A row of table for apple

Discussion

Riley, K. (IBPGR) : Can the documentation system of NIAR be utilized in gene banks of other countries?

Answer : Yes but with some modifications.

Riley, K. (IBPGR) : Comment : The challenge is both to develop powerful computer systems that can manage the huge information needed to keep track of nucleotide sequences in genome of plants as well as to develop information systems appropriate for PGR management in smaller national programs with limited resources.